

A Note on Choosing the Threshold for Large Covariance Estimations in Factor Models

Yuan Liao*

Rutgers University

August 31, 2016

Abstract

This note shows that for i.i.d. data, estimating large covariance matrices in factor models can be casted using a simple plug-in method to choose the threshold:

$$\mu_{jl} = \frac{c_0}{\sqrt{n}} \Phi^{-1}\left(1 - \frac{\alpha}{2p^2}\right) \sqrt{\frac{1}{n} \sum_{i=1}^n \hat{u}_{ji}^2 \hat{u}_{li}^2}.$$

This is motivated by the tuning parameter suggested by Belloni et al. (2012) in the lasso literature. It also leads to the minimax rate of convergence of the large covariance matrix estimator. Previously, the minimaxity is achievable only when $n = o(p \log p)$ by Fan et al. (2013), and now this condition is weakened to $n = o(p^2 \log p)$. Here n denotes the sample size and p denotes the dimension.

1 Introduction

Estimating large covariance matrices in factor models has been an important research area in recent years. This note provides a practical guidance on how to choose the threshold value for the method introduced in Fan et al. (2013). Consider a simple factor model:

$$y_{ji} = \lambda_j' f_i + u_{ji}, \quad i \leq n, j \leq p,$$

where λ_j denotes a K -dimensional vector of loadings for the j th individual, and f_i is a vector of common factors for the i th observation. In this model, only y_{ji} is observable. In the usual notation for factor models, the dependent variable is often denoted by y_{it} for the i the individual observed at time t . In this note, however, we shall apply the moderate

*Department of Economics, Rutgers University, New Brunswick, NJ 08901.

deviation theory for self-normalized sequences of independent data (de la Pena et al., 2009). Therefore, we stick to the more traditional notation in the statistical literature, and use $j \leq p$ to denote the index of variables and $i \leq n$ as the observations. That being said, the serial correlation is ruled out, due to the technical tools we are using to approximate the distribution of self-normalized sums.

The object of interest is to estimate the $p \times p$ covariance matrix of $u_n = (u_{1n}, \dots, u_{pn})'$, denoted by Σ_u , from the observations $\{y_{ji}\}_{j \leq p, i \leq n}$. Estimating Σ_u leads to many interesting applications. First, it makes it possible to obtain a good estimate of the covariance for $y_i = (y_{1i}, \dots, y_{pi})'$. Secondly, it allows as to improve the estimation of factors and loadings in the presence of cross-sectional correlations (Choi, 2012; Bai and Liao, 2013). Third, it also “activates” many classical Wald statistics for high-dimensional testing problems, which otherwise cannot handle the difficulty of using a large inverse weight matrix (Fan et al., 2015). In the so-called “approximate factor models”, this is a large and non-diagonal covariance.

I now describe Fan et al. (2013)’s estimator. Let \hat{f}_i and $\hat{\lambda}_j$ respectively denote the factors and loading estimators, which can be obtained via, e.g., the principal components method. Then we obtain the residual estimate $\hat{u}_{ji} = y_{ji} - \hat{\lambda}_j' \hat{f}_i$. The residual sample covariance is then

$$\hat{S}_u = (\hat{s}_{jl})_{p \times p}, \quad \hat{s}_{jl} = \frac{1}{n} \sum_{i=1}^n \hat{u}_{ji} \hat{u}_{li}.$$

Next, we apply soft-thresholding to obtain $\hat{\Sigma}_u = (\hat{\sigma}_{jl})_{p \times p}$, where

$$\hat{\sigma}_{jl} = \begin{cases} \hat{s}_{jl}, & j = l \\ \text{sgn}(\hat{s}_{jl})(|\hat{s}_{jl}| - \mu_{jl})_+, & j \neq l. \end{cases}$$

Here $\text{sgn}(\hat{s}_{jl})$ denotes the sign of \hat{s}_{jl} , and $(x)_+ = \max(x, 0)$. What plays the central role is the user-specified thresholding value μ_{jl} , which may depend on (n, p) , but we suppress such dependence in the notation. More importantly, its dependence on (j, l) indicates that the threshold should not be chosen as a universe constant. Ideally, it should be the smallest constant that just dominates the statistical error $|\hat{s}_{jl} - Eu_{ji}u_{li}|$. Fan et al. (2013) took the correlation matrix as their standpoint, which left a constant unspecified in the thresholding value, and suggested choose it by the cross-validation. In addition, the choice of the constant also depends on the number of factors to use in the model.¹

¹I communicated with Michael Wolf on this procedure when applied to the portfolio selection problems. Michael implemented it on the daily returns of 252 trading days, with the number of stocks varying from 30 through 500, and suggested using the constant “one” with five factors.

This note suggests, in contrast, a simple plug-in choice for μ_{jl} as follows:

$$\mu_{jl} = \frac{c_0}{\sqrt{n}} \Phi^{-1}\left(1 - \frac{\alpha}{2p^2}\right) \sqrt{\frac{1}{n} \sum_{i=1}^n \hat{u}_{ji}^2 \hat{u}_{li}^2}. \quad (1.1)$$

Here $c_0 > 1$ is taken as a constant arbitrarily close to one, e.g., $c_0 = 1.1$. This is equivalent to thresholding the studentized \hat{s}_{jl} using a universe constant $c_0 \Phi^{-1}(1 - \frac{\alpha}{2p^2})$. Here Φ^{-1} denotes the inverse standard normal CDF; α is a small significant level so that

$$P(\max_{jl \leq p} |\hat{s}_{jl} - Eu_{ji}u_{li}|/\mu_{jl} > 1) \leq \alpha + o(1).$$

We can choose, e.g., $\alpha = 0.05$. Everything else in the definition of μ_{jl} is completely data-driven, and is easy to plug in. This method uses the fact that the distribution of the normalized average

$$|\frac{1}{n} \sum_{i=1}^n u_{ji}u_{li}| / \sqrt{\frac{1}{n} \sum_{i=1}^n u_{ji}^2 u_{li}^2}$$

can be well approximated by the standard normal distribution, and was previously studied by Cai and Liu (2011) for estimating sparse covariances. The idea was also used commonly in the high-dimensional lasso literature (Belloni et al. (2012, 2014)).

It is helpful to look at the components of \hat{s}_{jl} more carefully. In fact,

$$\hat{s}_{jl} - Eu_{ji}u_{li} = \underbrace{\frac{1}{n} \sum_{i=1}^n \hat{u}_{ji} \hat{u}_{li} - \frac{1}{n} \sum_{i=1}^n u_{ji} u_{li}}_{\text{estimating residuals}} + \underbrace{\frac{1}{n} \sum_{i=1}^n u_{ji} u_{li} - Eu_{ji}u_{li}}_{\text{empirical process}}.$$

The plug-in thresholding value in fact bounds the empirical process components uniformly over all the individuals. But it does not control the “estimating residuals” components. Using the Cauchy-Schwarz inequality, Fan et al. (2013) showed that this component has a rate of convergence $O_P(\frac{1}{\sqrt{p}} + \sqrt{\frac{\log p}{n}})$, hence cannot be ignored. In fact, we will see that this rate can be improved to

$$\max_{jl \leq p} |\frac{1}{n} \sum_{i=1}^n \hat{u}_{ji} \hat{u}_{li} - \frac{1}{n} \sum_{i=1}^n u_{ji} u_{li}| = O_P(\frac{\log p}{n} + \frac{1}{p}).$$

Therefore, it is negligible so long as $n = o(p^2 \log p)$.

On the other hand, when $n = o(p^2 \log p)$ is not satisfied, the plug-in choice would under-threshold the residual sample covariance. The amount of under-thresholding would be at

most $O_P(\frac{1}{p^2}) \times p^2$ when the rate of convergence is with respect to the squared frobenius norm.

In this note, we write $\|A\|_{\max} = \max_{ij} |A_{ij}|$, and $\|A\| = \sqrt{\nu_{\max}(A'A)}$, where $\nu_{\max}(A)$ denotes the maximum eigenvalue of A . We state all the propositions without providing their proofs. The technical proofs follow from standard arguments in this literature, and are available upon requests.

2 Identification

As we only observe y_{ji} , consistently estimating Σ_u in any reasonable sense is possible only if u_i can be approximately identified from the factor components. Write Y and U the $p \times n$ matrices of y_{ji} and u_{ji} . Write Λ as the $p \times K$ matrix of λ_j and F as the $n \times K$ matrix of f_i . Then the matrix form of the model is

$$Y = \Lambda F' + U.$$

This implies $YY' = \Lambda F' F \Lambda' + \Lambda F' U' + (\Lambda F' U')' + UU'$. Now take the expectation on both sides yields

$$\frac{1}{pn} E(YY') = \frac{1}{p} \Lambda E f_i f_i' \Lambda' + \frac{1}{pn} E(UU').$$

We now see that the ‘‘pervasive condition’’ plays a central role in the identification:

The minimum eigenvalue of $\Lambda' \Lambda$ dominates the maximum eigenvalue of Σ_u .

Further suppose all the eigenvalues of $E f_i f_i'$ are bounded away from zero. Given these conditions, as $p \rightarrow \infty$, the first K eigenvalues of $\frac{1}{p} \Lambda E f_i f_i' \Lambda'$ do not vanish, but all the eigenvalues of $\frac{1}{pn} E(UU')$ decays to zero.

In addition, right multiplying Λ yields

$$[\frac{1}{pn} E(YY') - \frac{1}{pn} E(UU')] \Lambda = \Lambda E f_i f_i' \frac{1}{p} \Lambda' \Lambda$$

Hence there is a $K \times K$ matrix H so that the columns of ΛH are the first K eigenvectors of $\frac{1}{pn} E(YY') - \frac{1}{pn} E(UU')$, which are then approximately the first K eigenvectors of $\frac{1}{pn} E(YY')$ since $\frac{1}{pn} E(UU')$ is dominated as $p \rightarrow \infty$, due to the sin-theta theorem. Hence, as the dimension diverges, the pervasive condition ensures that Λ can be identified up to a rotation matrix.

Now left multiplying $\frac{1}{p}H'\Lambda'$ on both sides of $Y = \Lambda F' + U$ yields:

$$H^{-1}F' = \left(\frac{1}{p}H'\Lambda'\Lambda H\right)^{-1}\frac{1}{p}H'\Lambda'Y - \left(\frac{1}{p}H'\Lambda'\Lambda H\right)^{-1}H'\frac{1}{p}\Lambda'U.$$

The second term on the right hand side is dominated under suitable conditions that $\frac{1}{p}\Lambda'U$ is “smaller” than $\frac{1}{p}\Lambda'\Lambda F'$. Therefore, the asymptotic identification of ΛH yields the asymptotic identification of $H^{-1}F'$, as well as that of

$$U = Y - \Lambda H H^{-1}F'.$$

The following theorem gives a formal identification result using the $\|\cdot\|_{\max}$ norm.

Proposition 2.1. *Let $\nu_1 \geq \dots \geq \nu_p$ be the eigenvalues of $\frac{1}{n}E(Y Y')$, and let ξ_1, \dots, ξ_p be the corresponding eigenvectors. Suppose all the eigenvalues of $\Lambda'\Lambda/p$ are bounded away from zero and infinity, and $\|\Sigma_u\| = O(1)$. Then*

$$\|\Sigma_u - \sum_{l=K+1}^p \nu_l \xi_l \xi_l'\|_{\max} = O\left(\frac{1}{\sqrt{p}}\right).$$

We omit the proof of this proposition in this note. Heuristically, the key step is to prove $\|\frac{1}{p}\Lambda E f_i f_i' \Lambda' - \sum_{l=1}^K \nu_l \xi_l \xi_l'\|_{\max} = O(\frac{1}{\sqrt{p}})$, by showing $\|\xi_l - (\Lambda H)_l \|(\Lambda H)_l\|^{-1}\| = O(p^{-1})$ and $|\nu_l - \|(\Lambda H)_l\|^2| = O(\|\Sigma_u\|)$ for some $K \times K$ matrix H and $l \leq K$, where $(\Lambda H)_l$ denotes the l th column of ΛH . Then this proposition follows from the decompositions $\frac{1}{n}E(Y Y') = \Lambda E f_i f_i' \Lambda' + \Sigma_u$, and $\frac{1}{n}E(Y Y') = \sum_{l=1}^K \nu_l \xi_l \xi_l' + \sum_{l=K+1}^p \nu_l \xi_l \xi_l'$.

3 Improved Rate of Convergence

We employ the PC estimator of Bai (2003); Fan et al. (2013) to estimate F, Λ and U . Recall that

$$\hat{s}_{jl} - E u_{ji} u_{li} = \underbrace{\frac{1}{n} \sum_{i=1}^n \hat{u}_{ji} \hat{u}_{li} - \frac{1}{n} \sum_{i=1}^n u_{ji} u_{li}}_{\text{estimating residuals}} + \underbrace{\frac{1}{n} \sum_{i=1}^n u_{ji} u_{li} - E u_{ji} u_{li}}_{\text{empirical process}}.$$

3.1 Empirical process

Using the moderate deviation theory for self-normalized sequences (de la Pena et al., 2009), the second difference on the right hand side is bounded by

$$\frac{1}{\sqrt{n}}\Phi^{-1}\left(1 - \frac{\alpha}{2p^2}\right)\sqrt{\frac{1}{n}\sum_{i=1}^n u_{ji}^2 u_{li}^2} \leq \mu_{jl}\frac{c_0 + 1}{2c_0}$$

with probability at least $1 - \alpha - o(1)$, uniformly in $j, l \leq p$. Here $c_0 > 1$ and μ_{jl} are defined in Introduction. Besides that the data are independent, the sufficient conditions are, for some $c, C > 0$,

$$\min_{j,l \leq p} E u_{ji}^2 u_{li}^2 > c, \quad \max_{j,l \leq p} E u_{ji}^6 < C, \quad (3.1)$$

and

$$\max_{j,l} \left| \frac{1}{n} \sum_{i=1}^n (u_{ji}^2 u_{li}^2 - \hat{u}_{ji}^2 \hat{u}_{li}^2) \right| = o_P(1). \quad (3.2)$$

In addition, it is required that $\log p = o(n^{1/3})$.

3.2 Estimating residuals

We provide additional regularity conditions.

Assumption 3.1. (i) $\{f_i, u_i\}_{i \leq n}$ are i.i.d. and sub-Gaussian.

(ii) $E(u_i | f_i) = 0$.

(iii) There is $C > 0$, $\max_{j \leq p} \|\lambda_j\| < C$.

(iv) All the eigenvalues of $\frac{1}{p}\Lambda'\Lambda$ are bounded away from both zero and infinity.

The following assumption requires the weak cross-sectional correlations.

Assumption 3.2. There is $C > 0$,

(i) $\max_{l \leq p} \frac{1}{np} \sum_{i=1}^n \sum_{j,m \leq p} |\text{cov}(u_{mi} u_{li}, u_{ji} u_{li})| < C$.

(ii) $\max_{l \leq p} \sum_{j=1}^p |E u_{ji} u_{li}| < C$.

(iii) $E\left(\frac{1}{\sqrt{p}} \sum_{j=1}^p u_{ji} u_{jk} - E u_{ji} u_{jk}\right)^4 < C$.

(iv) $E\left\|\frac{1}{\sqrt{p}} \sum_{j=1}^p \lambda_j u_{ji}\right\|^4 < C$.

Proposition 3.1. Under Assumptions 3.1, 3.2 and those of Proposition 2.1, we have

$$\max_{j,l \leq p} \left| \frac{1}{n} \sum_{i=1}^n \hat{u}_{ji} \hat{u}_{li} - \frac{1}{n} \sum_{i=1}^n u_{ji} u_{li} \right| = O_P\left(\frac{1}{p} + \frac{\log p}{n}\right).$$

We omit the proof of this proposition. Heuristically speaking, it can be shown by directly computing the expansions of the PC estimators $\hat{\lambda}_j$ and \hat{f}_i ,

$$\max_{j,l \leq p} \left| \frac{1}{n} \sum_{i=1}^n \hat{u}_{ji} \hat{u}_{li} - \frac{1}{n} \sum_{i=1}^n u_{ji} u_{li} \right| = O_P\left(\frac{\log p}{n}\right) + \max_{j \leq p} \left| \frac{1}{n} \sum_{i=1}^n (H'^{-1} f_i - \hat{f}_i) u_{ji} \right|$$

for some rotation matrix H . Importantly, Fan et al. (2013) used the Cauchy-Schwarz inequality to bound the second term on the right hand side, and reached a non-sharp rate of convergence: $O_P(\frac{1}{\sqrt{p}} + \frac{1}{\sqrt{n}})$. In fact, proposition 3.1 can be proved by showing that

$$\max_{j \leq p} \left| \frac{1}{n} \sum_{i=1}^n (H'^{-1} f_i - \hat{f}_i) u_{ji} \right| = O_P\left(\frac{1}{n} + \frac{1}{p} + \sqrt{\frac{\log p}{np}}\right).$$

For some $q \in [0, 1)$, define

$$m_p = \max_{j \leq p} \sum_{l=1}^p |Eu_{ji} u_{li}|^q.$$

Proposition 3.1 leads to two exciting results. One is an improved rate of convergence of the covariance matrix estimation. Note that the convergence rate of $\|\hat{\Sigma}_u - \Sigma_u\|$ is determined by that of $\max_{j,l \leq p} \left| \frac{1}{n} \sum_{i=1}^n \hat{u}_{ji} \hat{u}_{li} - Eu_{ji} u_{li} \right|$. Fan et al. (2013) showed that (in the special case $q = 0, m_p = O(1)$), $\|\hat{\Sigma}_u - \Sigma_u\| = O_P(\frac{1}{\sqrt{p}} + \sqrt{\frac{\log p}{n}})$. Due to Proposition 3.1, this rate can be sharpened to

$$\|\hat{\Sigma}_u - \Sigma_u\| = O_P\left(\frac{1}{p} + \sqrt{\frac{\log p}{n}}\right).$$

The impact of estimating the unknown factors, is therefore weakened to $O_P(\frac{1}{p})$. The other exciting implication is that it is now possible to use the simple plug-in choice μ_{jl} as the thresholding value, so long as $n = o(p^2 \log p)$. In this case $\max_{j,l \leq p} \left| \frac{1}{n} \sum_{i=1}^n \hat{u}_{ji} \hat{u}_{li} - \frac{1}{n} \sum_{i=1}^n u_{ji} u_{li} \right| = o_P(\sqrt{\frac{\log p}{n}})$, hence is a smaller order than the empirical process part. Consequently, with probability at least $1 - \alpha - o(1)$,

$$|\hat{s}_{jl} - Eu_{ji} u_{li}| \leq \mu_{jl}$$

uniformly in $j, l \leq p$. Hence for the sparse covariance Σ_u , most of the elements of the estimated sample residual covariance are dominated by μ_{jl} .

Corollary 3.1. *Use the thresholding value μ_{jl} given by (1.1). When $n = o(p^2 \log p)$ and*

$\log p = o(n^{1/3})$, under the assumptions of Proposition 3.1, and conditions (3.1) (3.2),

$$\|\widehat{\Sigma}_u - \Sigma_u\| = O_P(m_p(\frac{\log p}{n})^{(1-q)/2}).$$

This is the minimax convergence rate.

4 Conclusion

In this note, I sharpen the rate of convergence of $\max_{j,l \leq p} |\frac{1}{n} \sum_i \widehat{u}_{ji} \widehat{u}_{li} - E u_{ji} u_{li}|$ for the PC estimator in approximate factor models. The rate is faster than that of Fan et al. (2013) when $n = o(p^2 \log p)$. An immediate consequence is a simple plug-in choice of the thresholding value for the estimated idiosyncratic covariance matrix, which takes a similar type to that of Belloni et al. (2012) in the lasso literature. It also leads to the minimax rate of convergence of the large covariance matrix estimator. Previously, the minimaxity is achievable only when $n = o(p \log p)$ by Fan et al. (2013), and now this condition is weakened to $n = o(p^2 \log p)$.

References

- BAI, J. (2003). Inferential theory for factor models of large dimensions. *Econometrica* **71** 135–171.
- BAI, J. and LIAO, Y. (2013). Statistical inferences using large estimated covariances for panel data and factor models. Tech. rep., University of Maryland.
- BELLONI, A., CHEN, D., CHERNOZHUKOV, V. and HANSEN, C. (2012). Sparse models and methods for optimal instruments with an application to eminent domain. *Econometrica* **80** 2369–2429.
- BELLONI, A., CHERNOZHUKOV, V. and HANSEN, C. (2014). Inference on treatment effects after selection among high-dimensional controls. *The Review of Economic Studies* **81** 608–650.
- CAI, T. and LIU, W. (2011). Adaptive thresholding for sparse covariance matrix estimation. *Journal of the American Statistical Association* **106** 672–684.
- CHOI, I. (2012). Efficient estimation of factor models. *Econometric Theory* **28** 274–308.
- DE LA PENA, V. H., LAI, T. L. and SHAO, Q.-M. (2009). *Self-normalized processes. Probability and its Applications*. New York). Springer-Verlag, Berlin.
- FAN, J., LIAO, Y. and MINCHEVA, M. (2013). Large covariance estimation by thresholding principal orthogonal complements (with discussion). *Journal of the Royal Statistical Society, Series B* **75** 603–680.
- FAN, J., LIAO, Y. and YAO, J. (2015). Power enhancement in high-dimensional cross-sectional tests. *Econometrica* **83** 1497–1541.